Exploring Open Source Solutions in the Management of ETD Processes

CHETAN S SONAWANE
KMC COLLEGE, INDIA
Introduction

ETD being the most important research materials holds an importance. With the development in ICT, Theses and Dissertations (TD) started transforming in electronic form and came to be known as Electronic Theses and Dissertation (ETD). Virginia Tech (VT) has taken initiatives towards accepting TDs in electronic form from 1996. VT led the foundation for the establishment of NDLTD. The practice of maintaining TDs in both form i.e. Print and Electronic still continues in many countries. ETD repositories of developed and developing countries are mushrooming.
Open Access (OA) repositories will foster research also it will benefit to all those doing research, it will also avoid duplication. In many countries the importance of ETDs is well recognized and they’ve made mandatory submission of TDs in electronic form. The software's like DSpace, Eprints, Fedora, ETD-db, Islandora, Samvera (formerly Hydra) are widely used for hosting the repositories. The software’s lacks many important tools in the management of ETD processes. Add-on tools are created for this, some add-on solutions are proposed for implementations, these includes Discovery solutions, Artificial Intelligence and Recommender system.
Selection process of OSS for ETD Management

The management process of ETDs consist of steps like creation, ingestion, access, archiving, preservation and distribution. Open Source repository software provides facility for submission, ingestion, dissemination and retrieval. There are more than 15 OSSs available for managing ETD repositories. Proprietary solutions are also available for managing ETD repositories, but looking into cost involved these makes it ineligible to apply. OSS are available free of cost with capability to capture, store, index, preserve, and redistribute all scholarly research material in digital format.
The importance of OA is well recognized and for endorsing OA, the selection of OSS for managing ETD repositories make sense. For the implementation of OSS for managing ETD repositories, a manager need to ensure that sufficiency of experience of related staff in terms of programming and database management skills is there on selected platform. To achieve the best performance, the selected software should be compatible with Institution infrastructure and computer environment (such as Operating System platform, server type and security policies, etc). In order to select the most compatible software, each solution should be checked from the platform related aspect such as:
compatible operating systems, compatible databases, development language and hardware. One of the major challenges of OSS generally is its frequent updates and upgrading the old system is another big challenge, especially if it is managed by in-house technical manpower.
OSS offers many opportunities for development and customization, it also comes with many challenges, bug squashing is more challenging. With hosted repository platform, the library pays for customer support as part of the annual maintenance fees. As mentioned above there are more than 15 OSS available for managing ETD repositories. Some of the remarkable solutions are: DSpace, EPrints, Fedora, ETD-db, Islandora, Samvera, ABCD repository etc. DSpace, Eprints, Fedora and ETD-db are most widely used. Islandora, Samvera and ABCD ETD repositories are relatively new and emerging solutions which has caught attention.
In the following sections a review is presented of the software mentioned above:

**DSpace:** DSpace is an open source digital repository system that capture, store, indexes, preserve and redistribute an organization research output in digital form. DSpace has a comprehensive and customizable Dublin Core (DC) metadata collection for describing items individually. DSpace is compliant to OAI-PMH, and is written in Java with a choice of choosing either PostgreSQL or ORACLE database for storing data with two interfaces JSPUI and XMLUI. The interface can be be customized using
XML, and a standard organizational scheme. It also supports OAI-PMH v2.0 to export METS. The latest version of DSpace supports faceted search and browsing facility under Apache Solr. The DSpace code is currently licensed under BSD open source license.

**EPrints:** EPrints is an open source software suitable for creating digital repositories compliant with OAI-PMH with Perl based plug-in architecture for importing and exporting data and converting objects (for search engine indexing). EPrints is based on custom built repository and it support SWORD interchange protocol, meaning that digital object can be extracted and shared.
Fedora: Fedora is a robust, modular, open source repository system for the management and dissemination of digital content for access and preservation. The information can be accessed via web API and web based administrator GUI. It also support and provide RDF search, OAI-PMH, Gsearch (Fulltext) and rebuild utilities with multiple and customer driven front end. Fedora is written in Java and tested under Windows, Linux and Mac OS and distributed under Apache 2.0 open source license. Fedora provides a foundation upon which many types of repository framework can be built, including the popular Samvera and Islandora projects. DSpace and Fedora software is stewared by DuraSpace.
**ETD-db:** ETD-db is an open source software written using Perl script with MySQL database developed by VT. ETD-db uses ETD-MS metadata standard developed by NDLTD which is based on DC. ETD-db is endorsed by NDLTD and is available free to members of NDLTD. Despite this there is little directional development, unfortunately ETD-db has not had a major release since 2004. ETD-db lacks several features that are necessary for managing future ETD collection. In addition to the standard Perl installation, it is also necessary to install additional “Perl module” which enhances the functionality of the language, but for this expert professional is needed.
**Islandora:** Islandora is a new entrant in the Institutional Repository software market. Islandora is an open source software framework that focuses on collaborative management, and discovery of digital objects, and provides a comprehensive and interoperable starting point for the implementation of a graphical interface to a Fedora repository. Islandora ecosystem is comprised of three major components: Drupal, Islandora and Fedora. Drupal as front end of the system, offer an array of collaborative tool and application for presentation. Fedora, at the back end of the system, provide a data store within which digital asset are managed. Islandora act as a glue which hold the system together, facilitates the messaging services between the other parts of the ecosystem.
The core design principle is the separation of data from presentation in order to leverage the strength of individual component and make extensive use of other OSS tool. To communicate with Fedora, Islandora uses REST architecture to retrieve content from database in Fedora. To manipulate and prepare the data for presentation in Drupal, Islandora makes use of Drupal Hook System. Hooks are code snippet providing functionality for content model that contain information about data object. Any Drupal theme will automatically work with Islandora because it uses standard Drupal component. As the DuraSpace foundation has stewarded DSpace and Fedora, Software designed for DSpace works with Islandora.
**Samvera:** Samvera is an open source, multi-institutional project that gives institutions a framework to build and deploy robust and durable digital repositories (the “body”) supporting multiple “heads”: feature rich, digital asset management application with tailored workflows. Samvera provides both a technical solution and software as well as vibrant community infrastructure, giving like-minded institutions the ability to collaborate and realize the benefit of pooled development. Samvera software is free and open source and available under Apache 2 license. A Samvera repository solution provide functionality for the full range of CRUD services (Create, Read, Update and Delete). To do this it employs a number of Ruby on Rail...
based components (or “gems”) in conjunction with three other “best of breed” open source software product: Fedora, Solr, and Blacklight.

**ABCD repository:** ABCD software is fully based on ISIS technology, some of the technical characteristic which has made ABCD repository software suitable for ETD repository application are Freedom of database structure, ETD databases from the existing standard format, ETD database using CEPAL database for full text access, Full text indexing, the use of non Latin script, Integrated end user search etc.
Open Source ETD Submission Solutions

There are at least 5 open-source software solution for ETD submission management. They include TAPIR, Valet, OpenETD, Vireo, and Weller Project (formerly Jarrow). Some submission tools were developed as an add-on to the existing repository solutions. They were developed as a plug-in to the existing repository solution so as to integrate it into repository solutions. There were some years when no absolute activity took place with regard to development of open source ETD submission tool, but in recent years tools like OpenETD and Vireo were developed which provided
some kind of solutions, but there is still no software that can handle the entirety of submission process.

**TAPIR**: Thesis Alive Plug-in for Institutional Repositories or TAPIR which was developed by University of Edinburgh in 2003 was developed as a software package based on DSpace. It is available under a 3 clause “modified BSD license”. TAPIR was developed as an add-on tool for DSpace which was written in Java, however it was created long time back in 2003 and then after it did not provided new version of the software compatible with newer version of DSpace. It was created for DSpace 1.2.
Parts of the DSpace code were modified with TAPIR to deliver some of its core functionality. Currently DSpace 6.0 version is available for installation and TAPIR is unsuitable for it, this only makes TAPIR unlikely candidate for adoption.

**Valet:** Visionary Technology in Library Services Inc. (VTLS) which developed an open source ETD submission module named Valet and publicly released in 2005 is available under Mozilla Public license 1.1. Valet is a Perl based module but coupled with Fedora digital repository. There has been no update since from initial release in 2005, there is also very little documentation available which resulted in making the
installation and use difficult. ARROW (Australian Research Repositories Online to the World) Project which was released during the summer of 2008, led by Monash University in Australia forked Valet when they created a Java version called Squire. Squire release does not provide any kind of documentation nor does it got updated for the past nine years which makes it useless for adoption.

**OpenETD:** In 2010 Rutger University developed OpenETD software as an add-on solution as well as standalone submission system and is available under GNU General Public License 3.0. This system is repository agnostic.
The software export the Theses and Dissertation in METS/XML compliant files for import into the desired repository solution. OpenETD is based on Linux, MySQL, PHP, this ETD solution utilizes better web application architecture than its predecessor. It provide features like email notification, automatic margin and page number detection, along with simple use management facility, this system is quite compelling. OpenETD is based on post defense process. The software also offers very limited ability to customize forms or the logic around the submission process.
Vireo: Vireo is a Java based DSpace plug-in developed by Texas Digital Libraries, and is available under a 3 part “modified BSD license”. Vireo is an online system for the submission and management of ETDs. Vireo was released in 2011 and is under active development. The software was developed to integrate into DSpace, for ease of use and ability to make simple system customization has led to an active user community that is beginning to grow outside Texas area. Vireo offers many of the features as like that of OpenETD does, such as automatic email notification and templates. Although Vireo handles more of the process than any other system, beginning immediately after defense and handling correction in
addition to basic ingesting, it does not handle any pre-defense procedures, such as scheduling or initial committee approval.

**Weller ETD Project:** Weller ETD Project (formerly known as Jarrow) software was designed to integrate with Islandora, an institutional repository solution built on top of Fedora and Drupal. The Weller ETD Project strives to assist the scholar and institution in the publication and dissemination of their work. This ETD module was designed to serve two complementary goals: to collect information about the theses from the people involved and to manage the process preparing, approving, submitting, accepting, and publishing the theses, dissertation or projects.
The workflow management system is broken into three components, which work together to allow for widely different workflow to be configured. The three component are Data Model, form editor and the Workflow manager. Each component is dependent on the component before it. The Weller ETD Project began in early 2012 at University of North British Columbia (UNBC) as a joint initiative of the Library and the office of Dean of Graduate Programs. The Project was last modified in 2013, and is actively maintained and is under active development.
NDLTD developed an ETD specific item-level metadata schema ETD-MS, based on DC, is a primarily descriptive schema; it does not include element for life cycle management event or structural relationship. Most of the existing metadata standard fail to provide for the unique properties of ETDs. Most of the metadata standards have fail to provide the level of specificity required to sufficiently describe the item. ETD-MS published by NDLTD provide low barrier to entry and broad application, which focuses on repository interoperability.
Assigning appropriate metadata to ETDs can improve discoverability by increasing their visibility. To describe digital resources accurately, metadata creators try to follow, as closely as possible, the thinking of the creator/author and also to anticipate what users might want to discover and how they’ll retrieve the information. As noted by many researcher, one of key issues for information retrieval and all other context-based text management application is document indexing. The generation of accurate indexing term is fundamental to the discovery, use, and reuse of digital resources.

Most of the digital repository management software (e.g. DSpace, Eprints, Fedora, and Islandora) supports simple DC and Qualified DC by default.
“Simple Dublin Core” is a DC metadata that uses no qualifiers. It applies only 15 elements without any qualifiers. On the other hand “Qualified Dublin Core” uses additional qualifiers to increase specificity or precision of the metadata. The Dublin Core Metadata Element Set (DCMES) or DC consist of small set of resource description categories which is notably different from many of the other metadata schemes due to its ease of use and interoperability. As DCMES consist of 15 basic elements only it has several limitations. These field, being generic for any kind of digital resource, do not capture any specific information about specialized content such as maps, images, video, objects, learning materials, etds etc.
A generic metadata schema is not sufficient enough to describe different types of resources with all relevant elements. Along with providing more option for descriptive metadata, repository should support preservation, structural and right management. DSpace is most widely used repository solution, but it currently provide a limited amount of preservation metadata in the form of file type identification and a checksum for each bit stream submitted. There is also provenance data captured for each event such as submission, approval and edited metadata. All of this data is created and captured by DSpace itself, but there is a
strong need to make use of additional use-supplied metadata. Preservation metadata enhances the library’s ability to manage activities related to digital item’s format, authenticity, and stability over time.
Add-on solutions for ETDs

Some of the add-on solutions which can be proved useful for ETDs and should be considered for adoption include Discovery System, Artificial Intelligence (AI) and Recommender System. Discovery Service or Web Scale Discovery Service is now common in higher academic and research libraries. Discovery service will enable library users in discovering local ETDs and Global ETDs from a single search box and make library research as intuitive as Google. Many OSSs are available for providing Discovery Service. AI relevance for the retrieval of most useful documents by means of Artificial Intelligence Markup Language (AIML)
has been proved helpful. AI can help the library patrons in retrieving the most relevant documents. Another OSS proposed for Open Source ETD repositories is Recommender System. Recommender System are Software tools which provide suggestion for items that a user may be interested in. The most suitable Recommender System identified for ETD repository is Apache Mahout.
A discovery interface includes features such as relevancy-based search result, faceted navigation and other features consistent with web based resources. VuFind is widely adopted and most suitable Discovery solution for adoption. VuFind is based on AMP architecture and uses Apache Solr as its default text retrieval engine. It support OAI-PMH version 2.0 as its standard for harvesting and includes support for MARC and XML format. Apache Tika could help in refining the search results, its toolkit which extract and extend metadata and text from over a thousand different
file types (such as ppt, xls and pdf). All of these file types can be parsed through a single interface, which makes Tika useful for search engine indexing, content analysis, translation, and much more. VuFind system offer support for loading record from specific repository system like DSpace, Eprint and Greenstone. In Samvera repository Blacklight discovery tool is integrated. The globally available repositories can also be integrated with local repository solution for wider dissemination of research by means of discovery tools.
Artificial Intelligence (AI)

AI is still in its nascent stage and the time to adoption horizon expected is next four to five years. AI encompasses 5 broad general areas of research which includes (1) Expert System (2) Natural Language Processing (3) Fuzzy Logic (4) Robotics and (5) Neural networks. The AIML language which is a set of XML element capable of representing and linking natural language expression, allows the creation of engine capable of maintaining simple dialogue. AI agent such as chatterbots and location base services are shifting the
focus of librarian, while also optimizing the search engine result to increase student access. The AIML language which is an open source can help in enabling the use in research for improvement in creation of robotic engine able to recognize information written in natural languages. AIML is XML compliant language which makes it possible to customize an Alicebot or creating one from scratch from beginning, which can be created for DSpace using XML. Development of metadata parser software of digital repositories for DSpace platform which will convert the input from XML to AIML is necessary. The Centre University Euripides of Marilia – UNIVEM, called Univem Aberto has created metadata parser software XML2AIML.
The DSpace platform enable the digital repository web server to receive request from external media, off the server. With request we are able to have access to all digital files that fit the information passed by the DC will an XML response that will be sent to XML2AIML software. XML2AIML software can do search request to the server of the repositories, the metadata already created for the standard DSpace platform will be used. AIML language has several limitation in creating robotic conversational engine, limiting the number of question these robotic engines can be asked.
The goal of Recommender System is to generate meaningful recommendation to a collection of users for items or products that might interest them. Some of the examples of users of Recommender System are Amazon and Netflix. Recommender system are increasingly used on web to help user find material relevant to their interest. Recommender system have been proved useful by commercial establishment to boost sale, there have been less recommenders used in research. From a user point of view, recommenders are either personalized, i.e. recommendation
are based on knowledge of user’s preference or past activity, or non-personalized, recommendation are same for all users. There are two important classes of recommendation: collaborative filtering (CF) and content-based filtering (CBF). CF makes use of past interaction data to recommend new known item on to a user based on assumption of similarity (e.g. to other users in the case of CF). Another approach to build recommender system for repositories is to make use of CBF. This method attempt to recommend items using the assumption of content similarity, such as cosine, based on attribute (feature) of each item. Hybrid method
combines both CF and CBF. CF can not recommend any item no one has referred and is known as cold-item. The integration of recommender system in EPrints repository is easy as it is provided on EPrint Bazaar Plugin (http://bazaar.eprints.org/466) and is based on CBF. For DSpace Apache Mahout an OSS is an ideal solution based on CF. Although there is an alternative of CBF, CF recommenders are known for their better results in most cases. Hybrid approach can be applied by combining the results produced by CF and CBF and then after merging its results to produce a final list.
Although libraries invest huge amounts in building their collections and platform to host various resources including ETDs, sometimes these resources get under-utilized due to lack of powerful search tools and repository solutions inability in retrieving useful resources. Some of the solutions developed by Institutions/universities to support ETD management have not updated themselves on a regular basis and thus became irrelevant for adoption. Where as the add on solutions mentioned in this paper are created by commercial
establishment and are regularly updating and providing innovative approaches and hence should be considered for integration into ETD repository.
Thank You

Questions ???